

FORECASTING OVERDISPERSED COUNT TIME SERIES WITH THE DLD-INAR(1) MODEL.

TÉGAWENDÉ MARTIN KABORE^{1*} AND O. JEAN-ETIENNE OUEDRAOGO²

ABSTRACT. This paper investigates coherent forecasting of overdispersed count data using the stationary DLD-INAR(1) process. Although the process has been introduced previously, we extend the existing literature by establishing its k -step-ahead predictive properties and the corresponding predictive probability distribution. Based on this, we develop forecasting procedures using median and mode predictors, with parameters estimated via conditional maximum likelihood. The main contribution of this manuscript lies in the comprehensive evaluation of forecasting performance. Through simulation studies using PRMSE, PMAE, and PTP metrics, we show that forecasts based on the median and mode outperform those based on the mean. Finally, we demonstrate the practical relevance of our results through a real data application, comparing the predictive accuracy of the DLD-INAR(1) process with that of the GINAR(1) and PINAR(1) models.

Keywords. Stationary DLD-INAR(1) process, Coherent forecast, Overdispersion, Conditional Maximum Likelihood.

2020 Mathematics Subject Classification. 60G10, 62M10

1. INTRODUCTION

The analysis and forecasting of integer-valued time series have become essential steps for decision-making across various fields, including public health, economics, finance, computer science, and epidemiology. These series often exhibit mixed behavior due to the influence of multiple probabilistic mechanisms and frequently present overdispersion, where the variance exceeds the mean, as well as strong correlation between observations. Examples include the number of requests received by a web server per minute or per hour in computer science, and the number of people infected with poliomyelitis in epidemiology. To provide more accurate analysis and coherent forecasting of such series, Integer-Valued Autoregressive (INAR) processes have emerged as a preferred approach. Among these, the stationary DLD-INAR(1) process appears particularly well-suited to address the aforementioned characteristics. Introduced by T. M. Kaboré and co-authors [13], this stationary process is designed for integer-valued time series

Date: Received: Aug 20, 2025; Accepted: Sep 22, 2025.

* Corresponding author

© The Author(s) 2025. This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. To view a copy of the licence, visit <https://creativecommons.org/licenses/by-nc-nd/4.0/>.

whose partial autocorrelation function (PACF) suggests an order of one. It is constructed using the binomial thinning operator proposed by Van Harn [8] and relies on the discrete Lindley distribution as its marginal, which belongs to the family of discrete self-decomposable distributions (DSD). Forecasting count time series remains a methodological challenge due to the discrete nature of the data. A coherent forecasting approach, first proposed by Freeland and McCabe [12] for the Poisson-INAR(1) model, ensures integer-valued predictions by relying on the median and mode of the predictive distribution rather than the conditional mean, which may yield non-integer values. This methodology was later extended to INAR(2) models by Jung and Tremayne [5] and applied to binomial autoregressive processes by Kim and Park [6]. It was further developed by Maiti and Biswas [9] for the GINAR(1) process and more recently generalized by Awale et al. [7] to the NBINAR(1) process, demonstrating its effectiveness for large-scale and highly overdispersed time series. In this study, we focus on evaluating the forecasting performance of the DLD-INAR(1) process. Parameters for all models are estimated via Conditional Maximum Likelihood (CML), and the properties of the DLD-INAR(1) predictive distribution are analyzed. The primary objective is to comprehensively assess its forecasting capabilities by comparing its performance with that of the GINAR(1) and PINAR(1) models, while highlighting the superiority of median- and mode-based predictors over the mean predictor using key metrics such as Prediction Root Mean Squared Error (PRMSE), Prediction Mean Absolute Error (PMAE), and Percentage of True Prediction (PTP). This study thereby demonstrates the ability of the DLD-INAR(1) process to provide accurate and reliable forecasts for overdispersed integer-valued time series.

The paper is structured as follows. Section 2 introduces the stationary DLD-INAR(1) process and its conditional properties. Coherent forecasting is developed in Section 3 through the derivation of the predictive distribution. Section 4 presents a simulation study to assess the model's performance using PRMSE, PMAE, and PTP metrics. Subsequently, real-world count data applications are conducted in Section 5. Finally, Section 6 concludes the paper with key findings and remarks.

2. STATIONARY DLD-INAR(1) PROCESS

In this section, the focus will primarily be on presenting the stationary first-order autoregressive integer-valued process with a Discrete-Lindley Distribution (DLD) as its marginal distribution. In the first subsection, we will introduce the DLD distribution and characterize the distribution of the innovation term for the stationary DLD-INAR(1) process. In the second subsection, we will present the conditional measures of the process. Also, it should be noted that in this section we only present the various results of the DLD-INAR(1) process without providing the proofs, which are available in [13].

2.1. Description of the stationary DLD-INAR(1) process. The stationary first-order integer-valued auto-regressive process with discrete Lindley distribution marginal, denoted by DLD-INAR(1), was introduced by T. M. Kaboré [13], and is defined as

$$X_t = \alpha \star X_{t-1} + \epsilon_t, \quad \alpha \in (0, 1), t \geq 1, \quad (2.1)$$

where:

- (1) the marginal distribution of $\{X_t\}$ is DLD;
- (2) $\{X_t\}$ is independent of ϵ_t ;
- (3) $\alpha \star X_{t-1} = \sum_{i=1}^{X_{t-1}} W_i$, with $\{W_i\}$ as an i.i.d. sequence of Bernoulli(α) random variables, and consequently $\alpha \star X_{t-1} \mid X_{t-1} \sim \text{Bin}(X_{t-1}, \alpha)$.

Thus, the key characteristics of the DLD distribution, namely its probability mass function (pmf) and its generating function (gf), are defined as follows:

$$\begin{aligned} P(X = x) &= (e^\theta - 1)^2(1 + x)e^{-\theta(2+x)}, \quad \theta > 0, \quad x = 0, 1, \dots; \\ \phi_X(s) &= \left(\frac{e^\theta - 1}{e^\theta - s} \right)^2, \quad \theta > 0, \quad s \neq e^\theta. \end{aligned}$$

If this is to be the marginal distribution of the above process, then the innovation process ϵ_t must have the probability generating function (PGF)

$$\phi_\epsilon(s) = \alpha^2 + (1 - \alpha^2) \left[\frac{2\alpha}{1 + \alpha} \frac{e^\theta - 1}{e^\theta - s} + \frac{1 - \alpha}{1 + \alpha} \frac{(e^\theta - 1)^2}{(e^\theta - s)^2} \right] \quad \theta > 0, \quad s \neq e^\theta.$$

Consequently, the probability mass function of the distribution of the innovation term, ϵ_t is given by:

$$\begin{aligned} P(\epsilon = x) &= \alpha^2 I_0(x) + (1 - \alpha^2) \left[\frac{2\alpha}{1 + \alpha} (1 - e^{-\theta}) (e^{-\theta})^x \right. \\ &\quad \left. + \frac{1 - \alpha}{1 + \alpha} (1 + x) (1 - e^{-\theta})^2 (e^{-\theta})^x \right], \quad x = 0, 1, \dots, \end{aligned}$$

where, $I_0(x)$ is the degenerate distribution function at zero.

Thus, the mean and the variance of ϵ_t are given by respectively:

- (1) $\mu_\epsilon = \frac{2(1 - \alpha)}{e^\theta - 1}$;
- (2) $\sigma_\epsilon^2 = \frac{2e^\theta(1 - \alpha^2) - 2\alpha(e^\theta - 1)(1 - \alpha)}{(e^\theta - 1)^2}$.

From the above, the DLD-INAR(1) process can be rewritten in the following form:

$$X_t = \alpha \star X_{t-1} + I_t H_t, \quad \alpha \in (0, 1), t \geq 1,$$

where:

- (1) $\{I_t\}$ is a sequence of independent and identically distributed (i.i.d.) random variables with a Bernoulli $(1 - \alpha^2)$ distribution, independent of $\{H_t\}$;
- (2) $\{H_t\}$ is a mixture of **Geo** $(1 - e^{-\theta})$ and **NB** $(2; 1 - e^{-\theta})$ of respective weights $\frac{2\alpha}{1 + \alpha}$ and $\frac{1 - \alpha}{1 + \alpha}$. Where **Geo** and **NB** denote the Geometric and Binomial Negative distribution, respectively.

2.2. Conditional properties of DLD-INAR(1). This subsection discusses some properties of the stationary DLD-INAR(1) process, primarily related to its conditional measures.

Thus, for X_t , a stationary DLD-INAR(1) process, the conditional mean and variance are defined, respectively, as follows:

$$E(X_t | X_{t-1}) = \alpha X_{t-1} + \frac{2(1 - \alpha)}{(e^\theta - 1)};$$

and,

$$Var(X_t | X_{t-1}) = \alpha(1 - \alpha)X_{t-1} + \frac{2e^\theta(1 - \alpha^2) - 2\alpha(e^\theta - 1)(1 - \alpha)}{(e^\theta - 1)^2}.$$

Also, by considering two consecutive observations X_t and X_{t-1} of the stationary DLD-INAR(1) process, their joint probability generating function (PGF) is given by:

$$G_{X_t, X_{t-1}}(s_1, s_2) = \left(\frac{e^\theta - 1}{e^\theta - s_2(1 - \alpha + \alpha s_1)} \right)^2 + (1 - \alpha^2) \left(\frac{e^\theta - 1 + \alpha(1 - s_1)}{e^\theta - s_1} \right)^2.$$

It is observed that $G_{X_t, X_{t-1}}(s_1, s_2)$ is not symmetric, therefore, the DLD-INAR(1) process is not time-reversible. Therefore, the time non-reversibility of the DLD-INAR(1) process highlights a dynamic where the past and the future are not symmetric.

As for the transition function of the stationary DLD-INAR(1) process, it is defined as:

$$\begin{aligned} P(X_t = x_t | X_{t-1} = x_{t-1}) &= \sum_{k=0}^{\min(x_{t-1}, x_t)} \binom{x_{t-1}}{k} \alpha^k (1 - \alpha)^{x_{t-1} - k} \left[\alpha^2 I_0(x_t - k) \right. \\ &+ (1 - \alpha) \times e^{-\theta(x_t - k)} \left(2\alpha(1 - e^{-\theta}) \right. \\ &\left. \left. + (1 - \alpha)(1 - e^{-\theta})^2(1 + x_t - k) \right) \right]. \end{aligned} \quad (2.2)$$

The conditional log-likelihood function $L(x_t, \lambda)$ of the stationary DLD-INAR(1) process, X_t , with parameter $\lambda = (\alpha, \theta)$, is given by:

$$\begin{aligned} L(x_t, \lambda) &= \sum_{t=2}^n \ln \left[\sum_{k=0}^{\min(x_{t-1}, x_t)} \binom{x_{t-1}}{k} \alpha^k (1 - \alpha)^{x_{t-1} - k} \left[\alpha^2 I_0(x_t - k) + (1 - \alpha) \right. \right. \\ &\left. \left. \times e^{-\theta(x_t - k)} \left(2\alpha(1 - e^{-\theta}) + (1 - \alpha)(1 - e^{-\theta})^2(1 + x_t - k) \right) \right] \right]. \end{aligned}$$

This log-likelihood function will be used in Sections 4 and 5 to estimate the parameters via the maximum likelihood method. The estimated values will then be used to assess the accuracy of the forecasts using appropriate indicators.

3. COHERENT FORECASTING

In INAR(1) models, the conditional mean is traditionally used for forecasting, although it may yield non-integer values that are inconsistent with the discrete nature of the data (Awale et al. [7]). To address this issue, Freeland and McCabe [12] propose the median, which minimizes the conditional absolute error. Maiti and Biswas [9], on the other hand, advocate using the mode as a coherent forecast.

3.1. h-step ahead forecasting distribution. In this subsection, we aim to determine the k -step ahead forecasting distribution. To do so, we will first express X_{t+k} as a function of X_t , and then we will study the conditional measures at k -step ahead of X_{t+k} .

Proof. Let $\{X_t\}$ be an INAR (1) process. Then, k -step ahead, X_t is defined by:

$$X_{t+k} = \alpha^k \star X_t + \sum_{j=0}^{k-1} \alpha^j \star \epsilon_{t+k-j}.$$

$$\begin{aligned} X_{t+k} &= \alpha \star X_{t+k-1} + \epsilon_{t+k} \\ &= \alpha \star (\alpha \star X_{t+k-2} + \epsilon_{t+k-1}) + \epsilon_{t+k} \\ &= \alpha^2 \star (\alpha \star X_{t+k-3} + \epsilon_{t+k-2}) + \alpha \star \epsilon_{t+k-1} + \epsilon_{t+k} \\ &= \alpha^3 \star (\alpha \star X_{t+k-4} + \epsilon_{t+k-3}) + \alpha^2 \star \epsilon_{t+k-2} + \alpha \star \epsilon_{t+k-1} + \epsilon_{t+k} \\ &\vdots \\ &\vdots \\ X_{t+k} &= \alpha^k \star X_t + \sum_{j=0}^{k-1} \alpha^j \star \epsilon_{t+k-j}. \end{aligned}$$

□

Based on Proposition 3.1, then k -step-ahead conditional probability mass function of INAR(1) process is defined by:

$$P(X_{t+k} = y \mid X_t = x) = \sum_{l=0}^{\min(y,x)} P(\alpha^k \star X_t = l) P\left(\sum_{j=0}^{k-1} \alpha^j \star \epsilon_{t+k-j} = y - l\right). \quad (3.1)$$

The initial results of the paper focus on the k -step-ahead conditional mean and k -step-ahead conditional variance, which are given by Lemmas 3.1 and 3.2, respectively.

Lemma 3.1. *Let X_t be a DLD-INAR(1) process. Then its k -step-ahead conditional mean is obtained as follows:*

$$E(X_{t+k} \mid X_t = x) = \alpha^k x + \frac{(1 - \alpha^k)}{1 - \alpha} \mu_\epsilon. \quad (3.2)$$

Proof.

$$\begin{aligned}
\mathbb{E}(X_{t+k} \mid X_t = x) &= \mathbb{E}\left(\alpha^k \star X_t + \sum_{j=0}^{k-1} \alpha^j \star \epsilon_{t+k-j} \mid X_t = x\right) \\
&= \mathbb{E}\left(\alpha^k \star X_t \mid X_t = x\right) + \mathbb{E}\left(\sum_{j=0}^{k-1} \alpha^j \star \epsilon_{t+k-j} \mid X_t = x\right) \\
&= \alpha^k x + \mathbb{E}\left(\sum_{j=0}^{k-1} \alpha^j \star \epsilon_{t+k-j}\right) \\
&= \alpha^k x + \sum_{j=0}^{k-1} \alpha^j \mathbb{E}(\epsilon_{t+k-j}) \\
\mathbb{E}(X_{t+k} \mid X_t = x) &= \alpha^k x + \frac{(1 - \alpha^k)}{1 - \alpha} \mu_\epsilon.
\end{aligned}$$

□

It is observed that $\lim_{k \rightarrow \infty} \mathbb{E}(X_{t+k} \mid X_t = x) = \mu_X = \frac{2}{(e^\theta - 1)}$. Thus, as the k -step-ahead conditional mean converges to the unconditional mean, it implies that, in the long run (as k becomes very large), the forecast based on past information becomes independent of previous observations. In other words, as the forecasting horizon extends, the forecast approaches a constant mean, no longer dependent on the initial conditions or past values of the process.

Lemma 3.2. *Let X_t be a DLD-INAR(1) process. Then its k -step-ahead conditional variance is obtained as follows:*

$$\text{Var}(X_{t+k} \mid X_t = x) = \alpha^k(1 - \alpha^k)x + \frac{(1 - \alpha^{2k})}{(1 - \alpha^2)} \sigma_\epsilon^2 + \frac{(1 - \alpha^k)(\alpha - \alpha^k)}{(1 - \alpha^2)} \mu_\epsilon. \quad (3.3)$$

Proof.

$$\begin{aligned}
\text{Var}(X_{t+k} \mid X_t = x) &= \text{Var}\left(\alpha^k \star X_t + \sum_{j=0}^{k-1} \alpha^j \star \epsilon_{t+k-j} \mid X_t = x\right) \\
&= \text{Var}(\alpha^k \star X_t \mid X_{t-1} = x) + \text{Var}\left(\sum_{j=0}^{k-1} \alpha^j \star \epsilon_{t+k-j}\right) \\
&= \alpha^k(1 - \alpha^k)x + \sum_{j=0}^{k-1} \alpha^{2j} \text{Var}(\epsilon_{t+k-j}) + \sum_{j=0}^{k-1} \alpha^j(1 - \alpha^j) \mathbb{E}(\epsilon_{t+k-j})
\end{aligned}$$

$$\begin{aligned}
\text{Var}(X_{t+k} | X_t = x) &= \alpha^k(1 - \alpha^k)x + \sum_{j=0}^{k-1} \alpha^{2j} \sigma_\epsilon^2 + \sum_{j=0}^{k-1} \alpha^j(1 - \alpha^j) \mu_\epsilon \\
&= \alpha^k(1 - \alpha^k)x + \sum_{j=0}^{k-1} \alpha^{2j} \left(\sigma_\epsilon^2 - \mathbb{E}(\epsilon_t) \right) + \sum_{j=0}^{k-1} \alpha^j \mu_\epsilon \\
&= \alpha^k(1 - \alpha^k)x + \frac{(1 - \alpha^{2k})}{(1 - \alpha^2)} \left(\sigma_\epsilon^2 - \mu_\epsilon \right) + \frac{(1 - \alpha^k)}{1 - \alpha} \mu_\epsilon \\
\text{Var}(X_{t+k} | X_t = x) &= \alpha^k(1 - \alpha^k)x + \frac{(1 - \alpha^{2k})}{(1 - \alpha^2)} \sigma_\epsilon^2 + \frac{(1 - \alpha^k)(\alpha - \alpha^k)}{(1 - \alpha^2)} \mu_\epsilon.
\end{aligned}$$

□

One has, $\lim_{k \rightarrow \infty} \text{Var}(X_{t+k} | X_t = x) = \frac{\sigma_\epsilon^2}{1 - \alpha^2} + \frac{\alpha \mu_\epsilon}{1 - \alpha^2} = \sigma_X^2 = \frac{2e^\theta}{(e^\theta - 1)^2}$. Thus, as the k -step-ahead conditional variance converges to the unconditional variance, it indicates that the uncertainty associated with the forecast diminishes in the long run. In other words, as k increases, the uncertainty related to the forecast becomes constant and independent of the process history, reflecting the overall variance of the process without considering short-term effects.

To obtain the forecasting distribution, specifically the k -steps-ahead conditional probability mass function, we will first determine the k -step-ahead conditional probability generating function, which uniquely determines the k -step-ahead conditional probability mass function.

Lemma 3.3. *Let X_t be a DLD-INAR(1) process. Then its k -step-ahead conditional probability generating function, $\phi_{X_{t+k}|X_t}(s)$ is obtained as follows:*

$$\phi_{X_{t+k}|X_t}(s) = (1 - \alpha^k + \alpha^k s)^{X_t} \left(\frac{e^\theta - 1 + \alpha^k - \alpha^k s}{e^\theta - s} \right)^2. \quad (3.4)$$

Proof.

$$\begin{aligned}
\phi_{X_{t+k}|X_t}(s) &= \mathbb{E} \left(s^{X_{t+k}} | X_t \right) \\
&= \mathbb{E} \left(\left(s^{\alpha^k \star X_t + \sum_{j=0}^{k-1} \alpha^j \star \epsilon_{t+k-j}} \right) | X_t \right) \\
&= \mathbb{E} \left(s^{\alpha^k \star X_t} | X_t \right) \mathbb{E} \left(s^{\sum_{j=0}^{k-1} \alpha^j \star \epsilon_{t+k-j}} \right) \\
&= \phi_{\alpha^k \star X_t | X_t}(s) \phi_{\sum_{j=0}^{k-1} \alpha^j \star \epsilon_{t+k-j}}(s).
\end{aligned}$$

Since $\alpha \star X_t = \sum_{i=0}^{X_t} W_i \sim \text{Bin}(X_t, \alpha)$, we have $(\alpha^k \star X_t \mid X_t) \sim \text{Bin}(X_t, \alpha^k)$, which leads to the generating function $\phi_{\alpha^k \star X_t \mid X_t}(s) = (1 - \alpha^k + \alpha^k s)^{X_t}$. Thus,

$$\begin{aligned}
\phi_{\sum_{j=0}^{k-1} \alpha^j \star \epsilon_{t+k-j}}(s) &= \mathbb{E} \left(s^{\sum_{j=0}^{k-1} \alpha^j \star \epsilon_{t+k-j}} \right) \\
&= \prod_{j=0}^{k-1} \mathbb{E} \left(s^{\alpha^j \star \epsilon_{t+k-j}} \right) \\
&= \prod_{j=0}^{k-1} \phi_{\alpha^j \star \epsilon_{t+k-j}}(s) \\
&= \prod_{j=0}^{k-1} \phi_{\alpha^j \star \epsilon_t}(s) \\
&= \prod_{j=0}^{k-1} \left(\frac{e^\theta - 1 + \alpha^{k-j} - \alpha^{k-j} s}{e^\theta - 1 + \alpha^{k-j-1} - \alpha^{k-j-1} s} \right)^2 \\
&= \left(\frac{e^\theta - 1 + \alpha^k - \alpha^k s}{e^\theta - s} \right)^2.
\end{aligned}$$

Hence the result! \square

It is observed that $\lim_{k \rightarrow \infty} \phi_{X_{t+k} \mid X_t}(s) = \left(\frac{e^\theta - 1}{e^\theta - s} \right)^2 = \phi_{X_t}(s)$. Then, the k -step ahead forecasting distribution of $X_{t+k} \mid X_t$ converges to the marginal distribution of X_t . Consequently, for higher order k , $(X_{t+k} \mid X_t) \sim \text{DLD}(\theta)$.

From Lemma 3.3, the h -step-ahead conditional probability generating function of the term $\sum_{j=0}^{k-1} \alpha^j \star \epsilon_{t+k-j}$ can be written as follows:

$$\begin{aligned}
\phi_{\sum_{j=0}^{k-1} \alpha^j \star \epsilon_{t+k-j}}(s) &= \left(\frac{e^\theta - 1 + \alpha^k - \alpha^k s}{e^\theta - s} \right)^2 \\
&= \alpha^{2k} + (1 - \alpha^{2k}) \left[\frac{2\alpha^k}{1 + \alpha^k} \frac{e^\theta - 1}{e^\theta - s} + \frac{1 - \alpha^k}{1 + \alpha^k} \frac{(e^\theta - 1)^2}{(e^\theta - s)^2} \right].
\end{aligned}$$

Lemma 3.4. *Let X_t be a DLD-INAR(1) process. Then its k -step-ahead conditional probability mass function, $P(X_{t+k} = x_{t+k} \mid X_t = x_t)$ is obtained as follows:*

$$\begin{aligned}
P(X_{t+k} = x_{t+k} \mid X_t = x_t) &= \sum_{l=0}^{\min(x_t, x_{t+k})} \binom{x_t}{l} \alpha^{kl} (1 - \alpha^k)^{x_t - l} \left[\alpha^{2k} I_0(x_{t+k} - l) \right. \\
&\quad + (1 - \alpha^k) \times e^{-\theta(x_{t+k} - l)} \left(2\alpha^k (1 - e^{-\theta}) \right. \\
&\quad \left. \left. + (1 - \alpha^k)(1 - e^{-\theta})^2 (1 + x_{t+k} - l) \right) \right],
\end{aligned}$$

where $I_0(x_{t+k} - l)$ is the degenerate distribution at zero.

Proof. From Equation (3.1), one has:

$$P(X_{t+k} = x_{t+k} \mid X_t = x_t) = \sum_{l=0}^{\min(x_t, x_{t+k})} P(\alpha^k \star x_t = l) P\left(\sum_{j=0}^{k-1} \alpha^j \star \varepsilon_{t+k-j} = x_{t+k} - l\right).$$

$$\text{However, } P(\alpha^k \star x_t = l) = \binom{x_t}{l} \alpha^{kl} (1 - \alpha^k)^{x_t - l},$$

because $(\alpha^k \star x_t \mid X_t) \sim \text{Bin}(x_t, \alpha^k)$,

and,

$$\begin{aligned} P\left(\sum_{j=0}^{k-1} \alpha^j \star \varepsilon_{t+k-j} = x_{t+k} - l\right) &= \frac{\phi_{\sum_{j=0}^{k-1} \alpha^j \star \varepsilon_{t+k-j}}(0)}{(x_{t+k} - l)!} \\ &= \alpha^{2k} I_0(x_{t+k} - l) + (1 - \alpha^k) \times e^{-\theta(x_{t+k} - l)} \left(2\alpha^k \right. \\ &\quad \left. \times (1 - e^{-\theta}) + (1 - \alpha^k)(1 - e^{-\theta})^2(1 + x_{t+k} - l)\right) \end{aligned}$$

one has:

$$\begin{aligned} P(X_{t+k} = x_{t+k} \mid X_t = x_t) &= \sum_{l=0}^{\min(x_t, x_{t+k})} \binom{x_t}{l} \alpha^{kl} (1 - \alpha^k)^{x_t - l} \left[\alpha^{2k} I_0(x_{t+k} - l) \right. \\ &\quad \left. + (1 - \alpha^k) \times e^{-\theta(x_{t+k} - l)} \left(2\alpha^k (1 - e^{-\theta}) \right. \right. \\ &\quad \left. \left. + (1 - \alpha^k)(1 - e^{-\theta})^2(1 + x_{t+k} - l)\right) \right]. \end{aligned}$$

□

3.2. 100(1 - γ)% prediction interval. Based on the k -step-ahead conditional probability mass function, otherwise called the k -step-ahead predictive probability distribution, given by Lemma 3.4, we can easily determine the 100(1 - γ)% prediction interval for X_{t+k} . Standard prediction intervals assume a symmetrically distributed predictive probability distribution. However, Figure 1 shows moderately positively skewed and unimodal predicted distributions. Thus, we determine the 100(1 - γ)% highest predictive probability (HPP) interval for X_{t+k} , defined as follows.

Definition 3.5. A 100(1 - γ)% highest predictive probability (HPP) interval $C_k = (X_L, X_U)$ for X_{t+k} given X_t is defined as

$$C_k = \{i : P(i \mid X_t) \geq h_\gamma\},$$

where h_γ is the largest number such that

$$P(X_L \leq X_{t+k} \leq X_U \mid X_t = x_t) = \sum_{i=X_L}^{X_U} P(i \mid x_t) \geq (1 - \gamma).$$

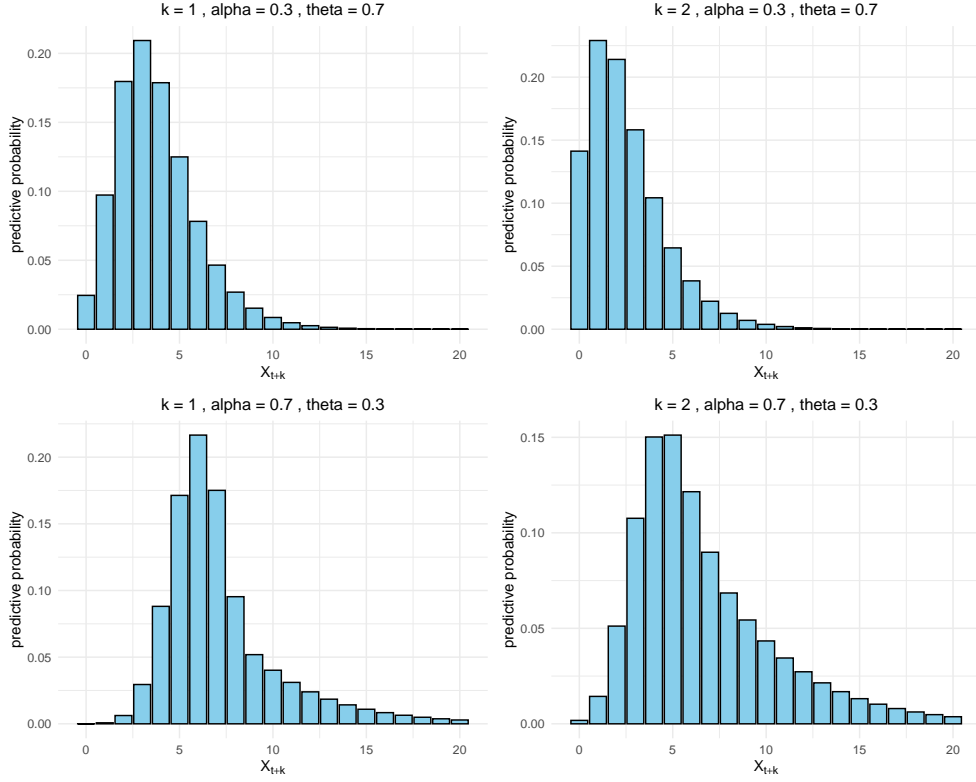


FIGURE 1. The k -step ahead predictive probability distribution for various parameter combinations: first row: $\alpha = 0.3, \theta = 0.7$; second row: $\alpha = 0.7, \theta = 0.3$.

3.3. Various measures of forecasting accuracy. Given an observed data set $X_1, \dots, X_n, X_{n+1}, \dots, X_{n+m}$ of size $n+m$, we partition the data into two sets. The training set, containing the first n observations, is used to estimate the parameters of the process. Based on the remaining m observations, called the test set, we define the following three descriptive measures of forecasting accuracy.

The first measure is the Prediction Root Mean Squared Error (denoted by PRMSE), defined as:

$$\text{PRMSE}(k) = \sqrt{\frac{1}{m} \sum_{i=1}^m (X_{n+i} - \hat{X}_{n+i}^{\text{me}})^2},$$

where $\hat{X}_{n+i}^{\text{me}} = \text{mean}(X_{n+i} \mid X_{n-k+i})$ is the k -step ahead conditional mean of the fitted process. The second measure is the Prediction Mean Absolute Error (denoted by PMAE), defined as:

$$\text{PMAE}(k) = \frac{1}{m} \sum_{i=1}^m |X_{n+i} - \hat{X}_{n+i}^{\text{med}}|,$$

where $\hat{X}_{n+i}^{\text{med}} = \text{median}(X_{n+i} \mid X_{n-k+i})$ is the k -step ahead conditional median of the fitted process. The third measure is the Percentage of True Predictions (denoted by PTP), which is defined as:

$$\text{PTP}(k) = \frac{1}{m} \sum_{i=1}^m I(X_{n+i} = \hat{X}_{n+i}) \times 100\%, \quad (3.5)$$

where $I(\cdot)$ is an indicator function, and predictions can be made through predictive mean, median, or mode. This measure gives us the number of true forecasts among 100 forecasts.

All these measures can be used to assess the forecasting accuracy of the models examined in the comparison study. Simulation results based on these forecasting measures are presented in the following section.

4. SIMULATION STUDY

This simulation study aims to assess the performance of the stationary DLD-INAR(1) process by analyzing various forecasting accuracy measures, including PRMSE, PMAE, PTP, and HPP prediction intervals. To this end, time series are generated according to the DLD-INAR(1) model using four parameter configurations defined as follows: $\alpha = 0.75, \theta = 1.5$; $\alpha = 0.75, \theta = 0.75$; $\alpha = 0.5, \theta = 1.5$; and $\alpha = 0.5, \theta = 0.75$.

The simulation is conducted in two phases. The first phase focuses on the analysis of the PRMSE, PMAE, and PTP indicators. For this purpose, a simulated dataset of 700 observations is generated according to the DLD-INAR(1) model, with 500 replications for each parameter configuration. This dataset is then split into two subsets:

- A training set of 400 observations, used for parameter estimation via the maximum likelihood method,
- A test set of 300 observations, used to assess predictive performance.

The evaluation based on the PRMSE and PMAE indicators aims to observe the behavior of these measures across different forecast horizons $h = 1, 2, 3, 4$, while the analysis of the PTP is intended to demonstrate the advantage of using the median and mode predictors over the mean predictor. The results for PRMSE and PMAE, presented in Table 1, show that errors increase with h , which is consistent with the growing uncertainty as the forecast horizon extends. Furthermore, the PTP results, reported in Table 2, indicate that the median and mode predictors outperform the mean predictor in forecasting the observations at each horizon h .

In the second phase of the study, training samples (size 400) and test samples (size 4) were used for different parameter sets. The $100(1 - \gamma)\%$ HPP intervals were computed from the training samples, and the predictions (mean, median, mode) were then evaluated on the test samples for various forecast horizons h . The results, based on 500 replications, are presented in Table 3, showing that the length of the HPP intervals increases with h , indicating the need to widen these intervals to maintain a constant coverage rate at horizon h .

TABLE 1. PRMSE and PMAE estimates for the simulated DLD-INAR(1) process.

h -step	PRMSE	PMAE	PRMSE	PMAE
	(a) $\alpha = 0.75, \theta = 1.5$		(b) $\alpha = 0.75, \theta = 0.75$	
1	0.557	0.238	1.255	0.652
2	0.809	0.424	1.545	0.944
3	0.915	0.540	1.787	1.182
4	0.949	0.594	1.875	1.310
	(c) $\alpha = 0.5, \theta = 1.5$		(d) $\alpha = 0.5, \theta = 0.75$	
1	0.890	0.518	1.595	1.092
2	0.947	0.610	1.792	1.310
3	1.004	0.646	1.814	1.380
4	1.119	0.678	2.012	1.518

TABLE 2. Estimated PTP using mean, median, and mode for the simulated DLD-INAR(1) process.

h -step	PTP			PTP		
	Mean	Median	Mode	Mean	Median	Mode
	(a) $\alpha = 0.75, \theta = 1.5$			(b) $\alpha = 0.75, \theta = 0.75$		
1	79.4	79.6	79.6	54.2	57.2	57.0
2	68.0	67.4	67.6	40.8	43.2	43.4
3	57.8	57.6	57.8	33.4	35.4	35.8
4	53.8	54.2	54.0	28.6	31.0	31.0
	(c) $\alpha = 0.5, \theta = 1.5$			(d) $\alpha = 0.5, \theta = 0.75$		
1	59.2	59.2	59.4	29.4	39.8	41.2
2	49.6	49.6	50.8	21.8	27.8	33.8
3	48.2	48.2	50.0	26.6	26.4	26.6
4	47.2	47.2	47.6	22.4	28.8	29.4

5. REAL-WORLD DATA ANALYSIS: POLIOMYELITIS DATASET

The objective of this real-data application phase is to fit the DLD-INAR(1) model to a monthly time series of poliomyelitis case counts reported by the United States Centers for Disease Control from 1970 to 1983, as documented by Zeger [3]. This dataset spans a fourteen-year period and comprises 168 observations. The results obtained using the DLD-INAR(1) model will be compared with those

TABLE 3. 95% HPP Prediction Intervals ($\gamma = 0.5$) for Simulations of the DLD-INAR(1) Model.

h -step	(Y_L, Y_U)	Mean	Median	Mode	(Y_L, Y_U)	Mean	Median	Mode
	(a) $\alpha = 0.75, \theta = 1.5$				(b) $\alpha = 0.75, \theta = 0.75$			
1	(0.03,1.51)	0.49	0	0	(0.39,4.18)	1.61	1	0
2	(0,1.77)	0.48	0	0	(0.14,4.79)	1.59	1	1
3	(0,2.06)	0.47	0	0	(0.04,5.03)	1.60	1	0
4	(0,2.15)	0.43	0	0	(0.01,5.17)	1.58	1	0
	(c) $\alpha = 0.5, \theta = 1.5$				(d) $\alpha = 0.5, \theta = 0.75$			
1	(0,2.01)	0.52	0	0	(0.08,4.98)	1.58	1	1
2	(0,2.15)	0.49	0	0	(0,5.29)	1.54	1	0
3	(0,2.16)	0.46	0	0	(0,5.32)	1.54	1	1
4	(0,2.17)	0.44	0	0	(0,5.33)	1.49	1	0

TABLE 4. Frequency distribution of the observed and expected monthly poliomyelitis cases from 1970 to 1983.

Number of cases	Observed	DLDINAR(1)	GINAR(1)	PINAR(1)
0	64	60	69	41
1	55	48	40	56
2	22	29	28	47
3	12	16	14	17
4	6	8	7	5
≥ 5	9	7	10	2
AIC		454.76	465.02	496.04

derived from the GINAR(1) and PINAR(1) models, as presented in [9], in order to assess the performance of the DLD-INAR(1) model relative to these two alternatives.

Our motivation for applying the DLD-INAR(1) model, designed to handle correlated and overdispersed count data, stems from several empirical observations. Figure 2, which displays the time series along with its autocorrelation function (ACF) and partial autocorrelation function (PACF), suggests that a first-order autoregressive (AR(1)) structure is appropriate for capturing the temporal dependence in the data. Furthermore, the series exhibits marked overdispersion, with a marginal mean of 1.333 and a variance of 3.505. Additionally, the frequency distribution shown in Table 4 reveals a moderate tail, indicating relatively high frequencies for larger count values. These characteristics strongly support the use of the DLD-INAR(1) model.

After fitting the three models to the series, Table 4 presents a comparison between the observed frequencies and the expected frequencies produced by each fitted model. It is evident that the DLD-INAR(1) model provides expected frequencies that are closer to the observed ones than those from the other two

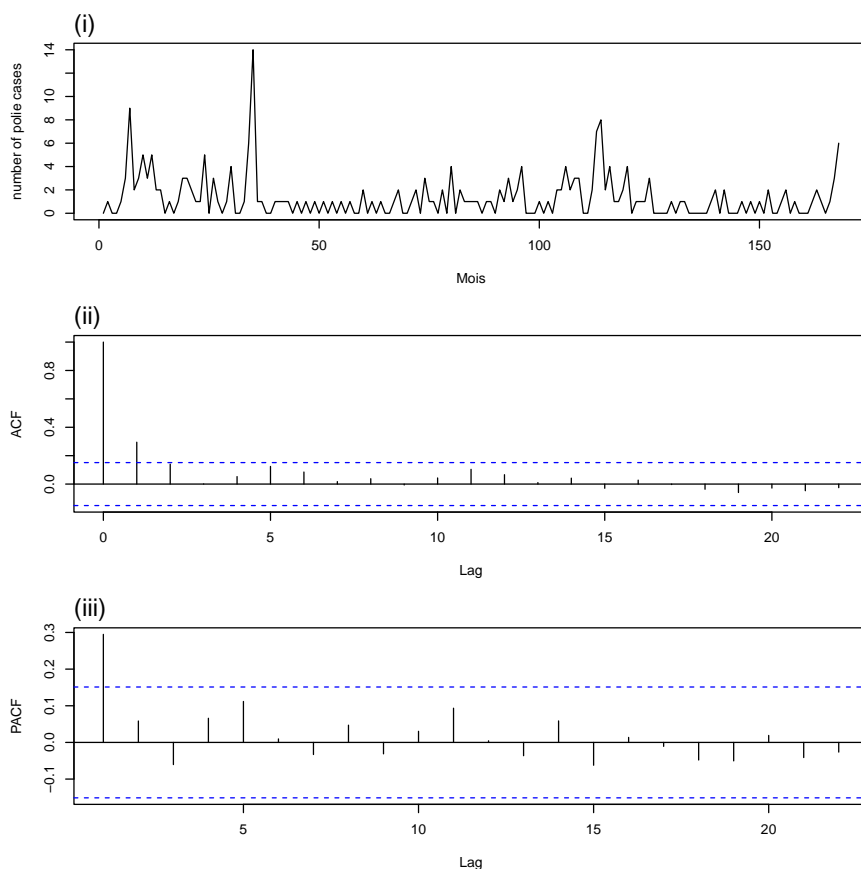


FIGURE 2. Poliomyelitis incidence data and its autocorrelation analysis: (i) Time series plot, (ii) ACF plot, (iii) PACF plot.

models, both in the tails and in the lower part of the distribution, indicating a superior overall fit.

As part of the model fitting procedure, we estimated the unknown parameters of all three models (DLD-INAR(1), GINAR(1), and PINAR(1)) using the conditional maximum likelihood estimation method and computed the corresponding Akaike Information Criterion (AIC) values. Subsequently, we conducted an h -step-ahead forecasting analysis based on three evaluation metrics: Predictive Root Mean Square Error (PRMSE), Predictive Mean Absolute Error (PMAE), and Predictive Turning Point (PTP), along with the corresponding Highest Predictive Probability (HPP) intervals. For this purpose, the time series was partitioned into two subsets: the training dataset, comprising the initial 138 observations, was utilized to estimate the parameters of the DLD-INAR(1), GINAR(1), and PINAR(1) models, while the test dataset, consisting of the remaining 30 observations, was employed to evaluate the forecasting performance metrics. The obtained results are presented in Tables 5 and 6.

Consistent with the results reported for the GINAR(1) and PINAR(1) models in [9], Table 5 clearly highlights the superior performance of the DLD-INAR(1)

TABLE 5. Parameter estimates, AIC, and PTP metrics based on the poliomyelitis data.

Model	Estimated values	AIC	h -step	PTP [median]	PTP [mode]
DLD-INAR(1)	$\hat{\alpha}_{\text{cml}} = 0.11$ $\hat{\theta}_{\text{cml}} = 0.87$	454.76	1	25	28
			2	24	27
			3	18	26
PINAR(1)	$\hat{\alpha}_{\text{cml}} = 0.32$ $\hat{\lambda}_{\text{cml}} = 0.97$	496.04	1	29.01	22.00
			2	29.08	30.51
			3	31.03	31.03
GINAR(1)	$\hat{\alpha}_{\text{cml}} = 0.32$ $\hat{\mu}_{\text{cml}} = 0.98$	465.02	1	28.00	41.00
			2	30.07	45.07
			3	34.00	48.00

process compared to the other models, both in terms of AIC and in forecasting the actual values when using the mode-based predictor. Specifically, the DLD-INAR(1) model yields the lowest AIC among the three models considered. Furthermore, for the DLD-INAR(1) process, the PTP values obtained using the mean-based predictor are relatively lower than those obtained with the mode-based predictor, clearly confirming the superior predictive performance of the latter. Additionally, the PTP values consistently decrease as the forecasting horizon increases, which is expected due to the growing uncertainty associated with longer-term predictions.

Table 6 displays the 98% HPP prediction intervals and the forecasts at various horizons h , obtained using the mean, median, and mode predictors. It is observed that for the DLD-INAR(1) process, the width of the HPP intervals increases with the horizon, which is generally expected. In contrast, this behavior is not observed for the GINAR(1) process, where the interval width remains nearly constant, and for the PINAR(1) process, where it even tends to decrease with the horizon. Moreover, the forecasts obtained from the mean, median, and mode predictors are identical.

Following the application to real data, the DLD-INAR(1) model proved to be superior to the GINAR(1) and PINAR(1) models. It achieved the lowest AIC and showed a better match between expected and observed frequency distributions. Additionally, in terms of PTP, it confirmed the superiority of the mode-based predictor over the mean-based one, while also respecting the expected decline in PTP values with increasing forecast horizon. Finally, it properly reflects the widening of HPP intervals as the horizon increases, which is consistent with the growing uncertainty associated with longer-term forecasts.

TABLE 6. 100(1- γ)% Prediction Intervals and Forecasts ($\gamma = 0.02$) for Poliomyelitis Data .

Model	h	Actual	(Y_L, Y_U)	Mean	Median	Mode
DLD-INAR(1)	1	0	(0,2)	1.53	1	1
	2	1	(0,2)	1.49	2	1
	3	0	(0,3)	1.37	1	0
	4	0	(0,3)	1.22	1	1
	5	0	(0,3)	1.22	1	0
GINAR(1)	1	0	(0,3)	1.51	1	1
	2	1	(0,3)	1.37	1	1
	3	0	(0,3)	1.32	1	1
	4	0	(0,3)	1.31	1	1
	5	0	(0,3)	1.30	1	1
PINAR(1)	1	0	(0,3)	1.51	1	1
	2	1	(0,2)	1.36	1	1
	3	0	(0,2)	1.32	1	1
	4	0	(0,2)	1.31	1	1
	5	0	(0,2)	1.30	1	1

6. CONCLUSION

In conclusion, this paper has advanced the study of coherent forecasting based on the median and mode of the predictive distribution for the DLD-INAR(1) process. Several key properties of k -step-ahead forecasts were highlighted, including the conditional mean and variance, as well as the conditional moment and probability generating functions, which fully characterize the predictive distribution. The Highest Predictive Probability (HPP) interval was defined as the interval with the highest predictive probability, and forecasting accuracy indicators PRMSE, PMAE, and PTP were used to confirm the superiority of median and mode predictions over the mean. The application to real data allowed a comparison of the DLD-INAR(1) process with the GINAR(1) and PINAR(1) models. The DLD-INAR(1) model consistently outperformed the others, which can be attributed to the specific properties of the Discrete Lindley distribution, particularly its ability to accommodate the observed overdispersion and tail behavior in the data, unlike the Geometric or Poisson distributions. As a perspective, an extension to order p appears as a natural approach to generalize the model and enhance its ability to capture more complex temporal dependencies.

ACKNOWLEDGMENTS

The authors wish to express their sincere gratitude to the editorial board and the reviewers for their careful reading of the article and for the constructive and insightful comments they provided.

REFERENCES

1. E. McKenzie, *Some simple models for discrete variate time series*, Water Resources Bulletin, **21**, 645–650 (1985). DOI: <https://doi.org/10.2307/1427362>
2. E. McKenzie, *Some ARMA models for dependent sequences of Poisson counts*, Advances in Applied Probability, **20**(4), 822–835 (1988). DOI: <https://doi.org/10.2307/1427362>
3. S. L. Zeger, *A regression model for time series of counts*, Biometrika, **75**(4), 621–629 (1988). DOI: <https://doi.org/10.1093/biomet/75.4.621>
4. M. A. Al-Osh, A. A. Alzaid, *First-order integer-valued autoregressive (INAR(1)) process*, Journal of Time Series Analysis, **8**(3), 261–275 (1987). DOI : <https://doi.org/10.1111/j.1467-9892.1987.tb00438.x>
5. Jung, R. C., & Tremayne, A. R. (2006). *Predictive methods for integer-valued autoregressive processes*. *Statistical Modelling*, **6**(4), 301-318. <https://doi.org/10.1191/1471082X06st115oa>
6. H. Kim, S. Park, *Coherent forecasting for binomial autoregressive processes*, Computational Statistics & Data Analysis, **54**(10), 2405–2417 (2010).DOI : <https://doi.org/10.5351/CKSS.2010.17.1.027>
7. M. Awale, T. V. Ramanathan, M. Kale, *Coherent forecasting in integer-valued AR(1) models with geometric marginals*, Journal of Data Science (2021).DOI:[https://doi.org/10.6339/JDS.201701_15\(1\).0006](https://doi.org/10.6339/JDS.201701_15(1).0006)
8. F. W. Steutel, K. van Harn, *Discrete analogues of self-decomposability and stability*, The Annals of Probability, **7**(5), 893–899 (1979). DOI: <https://doi.org/10.1214/aop/1176994950>
9. R. Maiti, A. Biswas, *Coherent forecasting for over-dispersed time series of counts data*, Brazilian Journal of Probability and Statistics, **29**(3), 529–550 (2015).DOI:<https://doi.org/10.1214/14-BJPS244>
10. M. M. Ristić, H. S. Bakouch, A. S. Nastić, *A new geometric first-order integer-valued autoregressive (NGINAR(1)) process*, Journal of Statistical Planning and Inference, **139**(7), 2218–2226 (2009). DOI:<https://doi.org/10.1016/j.jspi.2008.10.007>
11. M. M. Ristić, A. S. Nastić, H. S. Bakouch, *Estimation in an Integer-Valued Autoregressive Process with Negative Binomial Marginals (NBINAR(1))*, Communications in Statistics – Theory and Methods, **41**(19), 3622–3637 (2012). DOI:<https://doi.org/10.1080/03610926.2010.529528>
12. R. K. Freeland, B. P. M. McCabe, *Forecasting discrete valued low count time series*, International Journal of Forecasting, **20**(4), 427–434 (2004). DOI: [https://doi.org/10.1016/s0169-2070\(03\)00014-1](https://doi.org/10.1016/s0169-2070(03)00014-1)
13. T. M. Kabore, O. J.-E. Ouedraogo, M. Ilboudo, *First-order integer-valued autoregressive process with discrete Lindley distribution marginal*, Gulf Journal of Mathematics, **19**(1), 47–66 (2025). DOI: <https://doi.org/10.1214/aop/1176994950>
14. Gabr, M. M., Bakouch, H. S., & El-Taweel, H. M. (2025). *A first-order autoregressive process with weighted Lindley innovations and its applications to energy and financial data*. Annals of Mathematics and Computer Science, **29**, 1-19. <https://doi.org/10.56947/amcs.v29.573>

¹ L@MIA LABORATORY, NORBERT ZONGO UNIVERSITY, KOUDOUGOU BP376, BURKINA FASO.

Email address: tzorghoe@gmail.com

² L@MIA LABORATORY, NORBERT ZONGO UNIVERSITY, KOUDOUGOU BP376, BURKINA FASO.

Email address: ouedraogoetienne@yahoo.fr